(12) UK Patent Application (19) GB (11) 2 342 802 (13) A

(43) Date of A Publication 19.04.2000

(21) Application No 9916394.1

(22) Date of Filing 13.07.1999

(30) Priority Data
(31) 09173462 (32) 14.10.1998 (33) US

(51) INT CL[7]
H04N 7/15 , G06F 17/30

(52) UK CL (Edition R)
H4F FDX F32

(56) Documents Cited
EP 0660249 A1 WO 97/01932 A1 JP 060205151 A
US 5786814 A US 5729741 A US 5717869 A

(58) Field of Search
UK CL (Edition R) H4F FDX FEHM
INT CL[7] G06F 17/30 , G11B 27/34 , H04N 7/15
Online: WPI, EPODOC, JAPIO, INSPEC

(71) Applicant(s)
PictureTel Corporation
(Incorporated in USA - Delaware)
100 Minuteman Road, Andover,
Massachusetts 01810-1031, United States of America

(72) Inventor(s)
Steven L Potts
Peter L Chu

(74) Agent and/or Address for Service
Withers & Rogers
Goldings House, 2 Hays Lane, LONDON, SE1 2HW,
United Kingdom

(54) Abstract Title
Indexing conference content onto a timeline

(57) A method and system to index the content of conferences. It includes identifying each such conference participant producing a sound, capturing an image of each such conference participant, and with correlating the images of the conference participants with audio segments of an audio recording, that is the segments corresponding to the audio produced by the conference participant. The indexing system includes a sound recording mechanism, at least one identifier of locations of conference participants, a camera, an image storage device, a processor for associating the still images captured by the camera to the sound recorded by the sound recording mechanism thereby correlating still images of the conference participants to audio segments produced by the conference participants, and a graphical user interface which allows easy access to stored sound, images, and correlated data. It may also include an aiming device for pointing the camera at the person speaking.

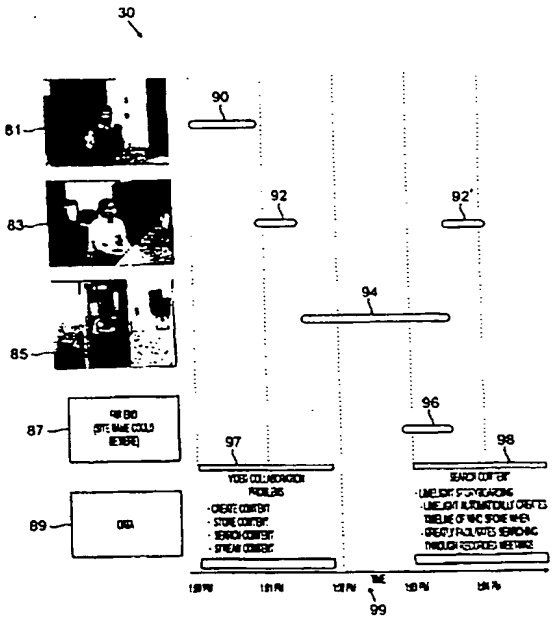

FIG. 3

GB 2 342 802 A

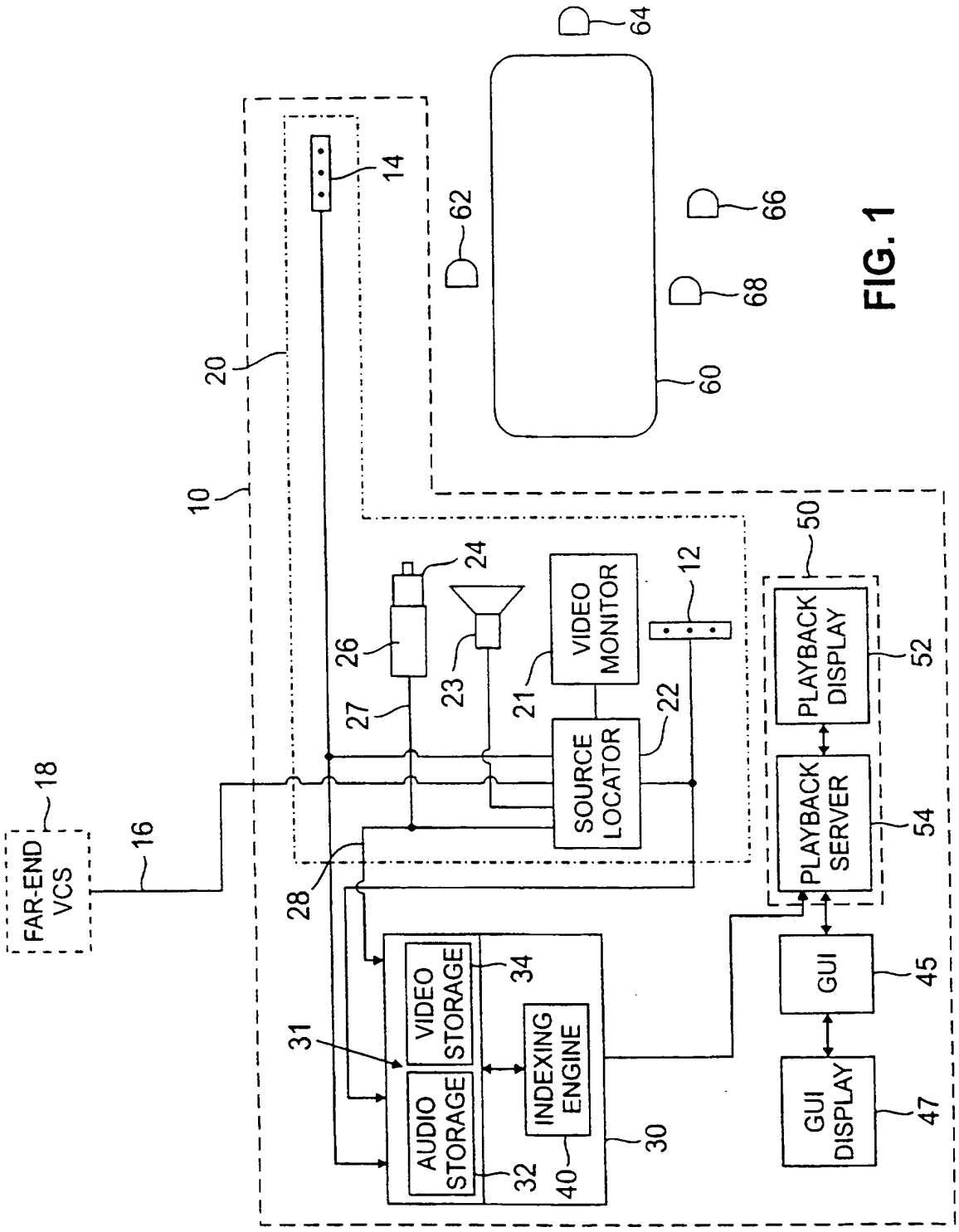
FAR-END VCS
18
16
10
20
14
62
64
66
68
60
24
26
27
23
21
VIDEO MONITOR
12
22
SOURCE LOCATOR
28
50
PLAYBACK DISPLAY
52
PLAYBACK SERVER
54
GUI
45
GUI DISPLAY
47
31
VIDEO STORAGE
34
AUDIO STORAGE
32
INDEXING ENGINE
40
30

FIG. 1

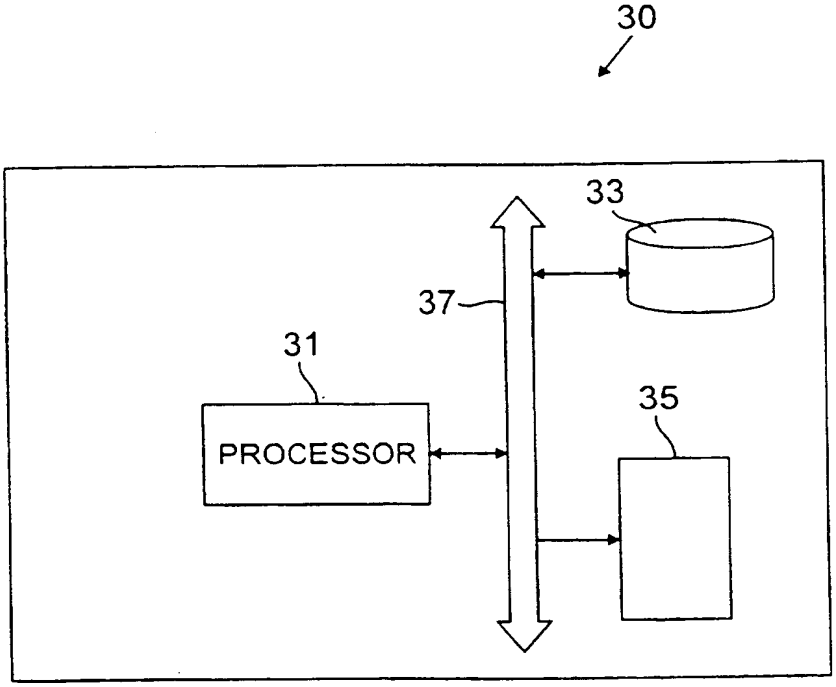

FIG. 2

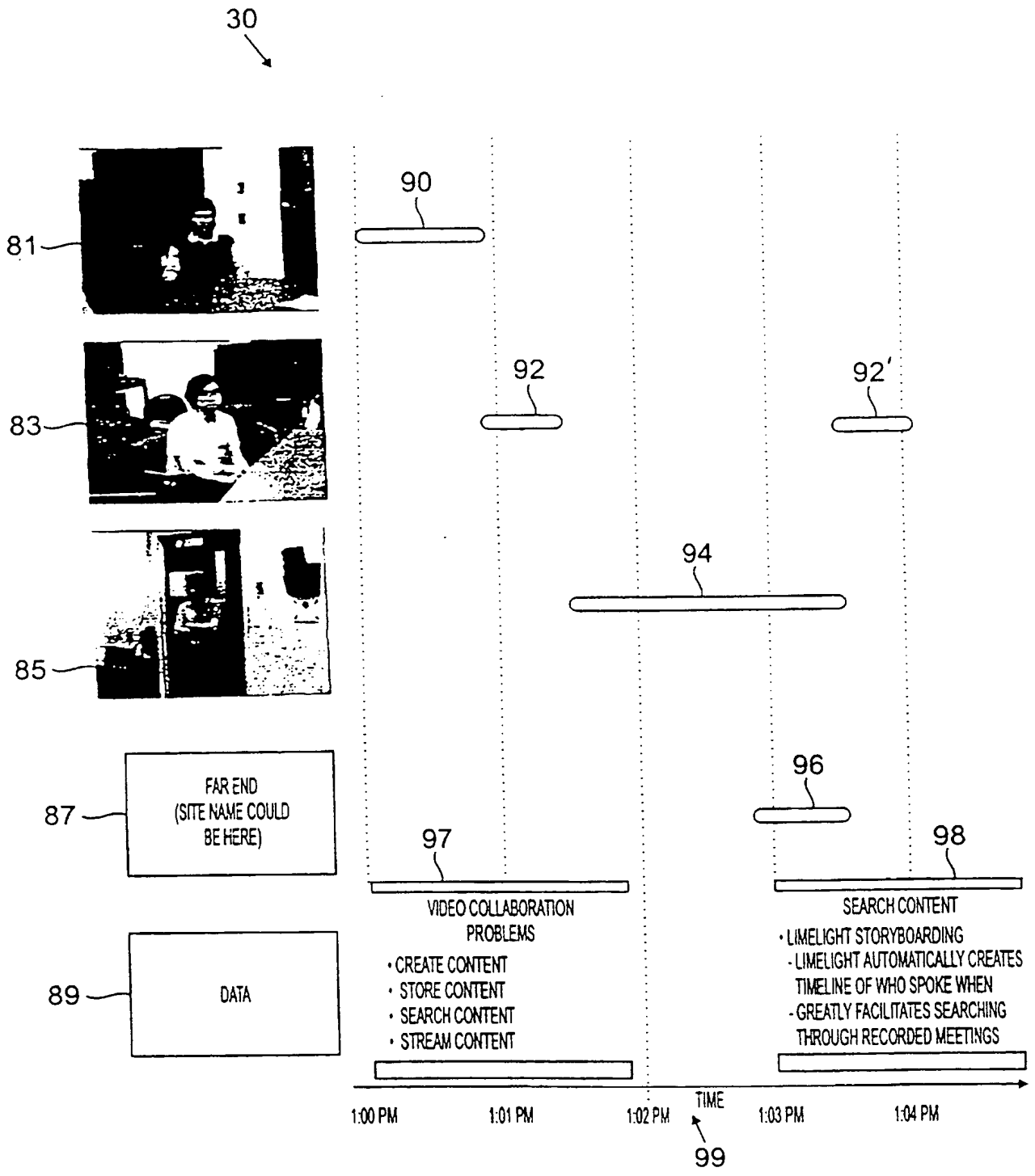
30
81
83
85
87
FAR END
(SITE NAME COULD
BE HERE)
89
DATA
90
92
92'
94
96
97
98
VIDEO COLLABORATION
PROBLEMS
• CREATE CONTENT
• STORE CONTENT
• SEARCH CONTENT
• STREAM CONTENT
SEARCH CONTENT
• LIMELIGHT STORYBOARDING
- LIMELIGHT AUTOMATICALLY CREATES
TIMELINE OF WHO SPOKE WHEN
- GREATLY FACILITATES SEARCHING
THROUGH RECORDED MEETINGS
1:00 PM
1:01 PM
1:02 PM
TIME
1:03 PM
1:04 PM
99

FIG. 3

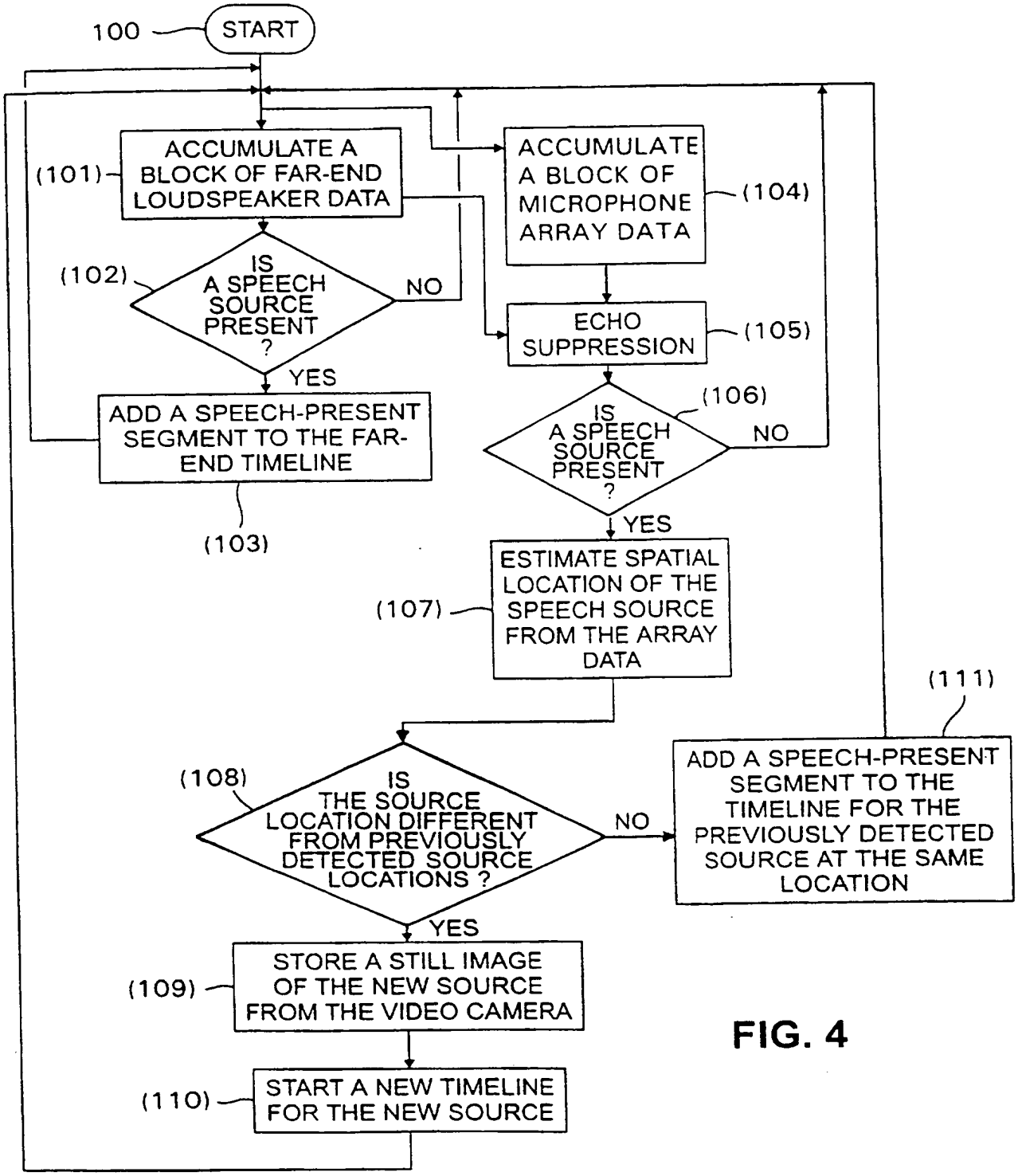
100
START
(101)
ACCUMULATE A BLOCK OF FAR-END LOUDSPEAKER DATA
(102)
IS A SPEECH SOURCE PRESENT ?
NO
YES
ADD A SPEECH-PRESENT SEGMENT TO THE FAR-END TIMELINE
(103)
ACCUMULATE A BLOCK OF MICROPHONE ARRAY DATA
(104)
ECHO SUPPRESSION
(105)
(106)
IS A SPEECH SOURCE PRESENT ?
NO
YES
(107)
ESTIMATE SPATIAL LOCATION OF THE SPEECH SOURCE FROM THE ARRAY DATA
(108)
IS THE SOURCE LOCATION DIFFERENT FROM PREVIOUSLY DETECTED SOURCE LOCATIONS ?
NO
YES
(109)
STORE A STILL IMAGE OF THE NEW SOURCE FROM THE VIDEO CAMERA
(110)
START A NEW TIMELINE FOR THE NEW SOURCE
(111)
ADD A SPEECH-PRESENT SEGMENT TO THE TIMELINE FOR THE PREVIOUSLY DETECTED SOURCE AT THE SAME LOCATION

FIG. 4

METHOD AND APPARATUS FOR INDEXING CONFERENCE CONTENT

This invention relates to the field of multimedia.

With the advent of economical digital storage media and sophisticated video/audio decompression technology capable of running on personal computers, thousands of hours of digitized video/audio data can be stored with virtually instantaneous random access. In order for this stored data to be utilized, it must be indexed efficiently in a manner allowing a user to find desired portions of the digitized video/audio data quickly.

For recorded conferences having a number of participants, indexing is generally performed on the basis of "who" said "what" and "when" (at what time). Currently used methods of indexing do not reliably give this information, primarily because video pattern recognition, speech recognition, and speaker identification techniques are unreliable technologies in the noisy, reverberant, uncontrolled environments in which conferences occur.

Also, a need exists for a substitute for tedious trial-and-error techniques for finding when a conference participant first starts speaking in a recording.

The invention features a method and a system for indexing the content of a conference by matching images captured during the conference to the recording of sounds produced by conference participants.

Using reliable sound source localization technology implemented with microphone arrays, the invention produces reliable information concerning "who" and "when" (which persons spoke at what time) for a conference. While

information concerning "what" (subject matter) is missing, the "who-when" information greatly facilitates manual annotation for the missing "what" information. In many search-retrieval situations, the "who-when" information alone will be sufficient for indexing.

In one aspect of the invention, the method includes identifying a conference participant producing a sound, capturing a still image of the conference participant, correlating the still image of the conference participant to the audio segments of the audio recording corresponding to the sound produced by the conference participant, and generating a timeline by creating a speech-present segment representing the correlated still image and associated audio segment. Thus, the timeline includes speech-present segments representing a still image and associated audio segments. The still image is a visual representation of the sound source producing the associated audio segments.

The audio recording can be segmented into audio segment portions and associated with conference participants, whose images are captured, for example, with a video camera.

Embodiments of this aspect of the invention may include one or more of the following features.

The still image of each conference participant producing a sound is captured as a segment of a continuous video recording of the conference, thereby establishing a complete visual indicator of all speakers participating in a conference.

The timeline is presented visually so that a user can quickly and easily access individual segments of the continuous recording of the conference.

The timeline can include a colored line or bar representing the duration of each speech segment with a

correlated image to index the recorded conference. The timeline can be presented as a graphical user interface (GUI), so that the user can use an input device (for example, a mouse) to select or highlight the appropriate part of the timeline corresponding to the start of the desired recording, access that part, and start playing the recording. Portions of the audio and video recordings can be played on a playback monitor.

Various approaches can be used to identify a conference participant. In one embodiment, a microphone array is used to locate the conference participant by sound. The microphone arrays together with reliable sound source localization technology reliably and accurately estimate the position and presence of sound sources in space.

The time elapsed from a start of the conference is stored with each audio segment and each still image. An indexing engine can be provided to generate the timeline by matching the elapsed time associated with an audio segment and a still image.

The system can be used to index a conference with only one participant. The timeline then includes an indication of the times in which sound was produced, as well as an image of the lone participant.

In applications in which more than one conference participant is present and identified, the system stores the times elapsed from the start of the conference and identifications of when a speaker begins speaking with each still image, a participant being associated with each image. The elapsed time is also stored with the audio recording each time a change in sound location is identified. The indexing engine creates an index, that is, a list of associated images and sound segments. Based on this index, a timeline is then generated for each still image (that is,

each conference participant) designating the times from the start of the conference when the participant speaks. The timeline also indicates any other conference participant who might also appear in the still image (for example, a neighbor sitting in close proximity to the speaker), but is silent at the particular elapsed time, thus giving a comprehensive overview of the sounds produced by all conference participants, as well as helping identify all persons present in the still images. The timeline may be generated either in real time or after the conference is finished.

In embodiments in which a video camera is used to capture still images of the conference participants, it can also be used to record a continuous video recording of the conference.

The system can be used for a conference with all participants in one room (near-end participants) as well as for a conference with participants (far-end participants) at another site.

Assuming that a speaker has limited movement during a conference, the same person is assumed to be talking every time sound is detected from a particular locality. Thus, if the speech source is determined to be the same as the locality of a previously detected conference participant, a speech-present segment is added to the timeline for the previously detected conference participant. If the location of a conference participant is different from a previously detected location of a near-end conference participant, a still image of the new near-end conference participant is stored and a new timeline is started for the new near-end conference participant.

In a video conference involving a far-end participant, the audio source is a loudspeaker at the near-

end transmitting a sound from a far-end speech source. The timeline is then associated with the far-end, and generating a timeline includes creating a speech-present segment for the far-end if a far-end speech source is present. Thus, a user of the invention can identify and access far-end speech segments. Further, if a far-end speech source is involved in the conference, echo can be suppressed by subtracting a block of accumulated far-end loudspeaker data from a block of accumulated near-end microphone array data.

Advantageously, therefore, a video image of a display presented at the conference is captured, and a timeline is generated for the captured video image of the display. This enables the indexing of presentation material as well as sounds produced by conference participants.

The present invention is illustrated in the following figures.

Fig. 1 is a schematic representation of a videoconferencing embodiment using two microphone arrays;

Fig. 2 is a block diagram of the computer which performs some of the functions illustrated in Fig. 1;

Fig. 3 is an exemplary display showing timelines generated during a videoconference; and

Fig. 4 is a flow diagram illustrating operation of the microphone array conference indexing method.

While the description which follows is associated with a videoconference between the local or near-end site and a distant or far-end site, the invention can be used with a single site conference as well.

Referring then to Fig. 1, a videoconference indexing system 10 (shown enclosed by dashed lines) is used to record and index a videoconference having, in this particular embodiment, four conference participants 62, 64, 66, 68 sitting around a table 60 and engaged in a videoconference. One or more far-end conference participants (not shown) also participate in the conference through the use of a local videoconferencing system 20 connected over a communication channel 16 to a far-end video conferencing system 18. The communication channel 16 connects the far-end video conferencing system to the near-end videoconferencing system 20 and far-end decompressed audio is available to a source locator 22.

Videoconference indexing system 10 includes videoconferencing system 20, a computer 30, and a playback system 50. Videoconferencing system 20 includes a display monitor 21 and a loudspeaker 23 for allowing the far-end conference participant to be seen and heard by conference participants 62, 64, 66, and 68. In an alternative embodiment, the embodiment shown in Fig. 1 is used to record a meeting not in a conference-call mode, so the need for the display monitor 21 and loudspeaker 23 of videoconferencing system 20 is eliminated. System 20 also includes microphone arrays 12, 14 for acquiring sound (for example, participants' speech), the source locator 22 for determining the location of a sound-producing conference participant, and a video camera 24 for capturing video images of the setting and participants as part of a continuous video recording. In one embodiment, source locator 22 is a stand-alone hardware, called "LIMELIGHT™", manufactured and sold by PictureTel Corporation, and which is a videoconferencing unit having an integrated motorized camera and microphone array. The "LIMELIGHT™" locator 22 has a digital signal

processing (DSP) integrated circuit which efficiently implements the source locator function, receiving electrical signals representing sound picked up in the room and outputting source location parameters. Further details of the structure and implementation of the "Limelight™" system is described in U.S. 5,778,082, the contents of which are incorporated herein by reference. (In other embodiments of the invention, multiple cameras and microphone configurations can be used.)

Alternative methods can be used to fulfill the function of source locator 22. For example, a camera video pattern recognition algorithm can be used to identify the location of an audio source, based on mouth movements. In another embodiment of the invention, an infrared motion detector can be used to identify an audio source location, for example to detect a speaker approaching a podium.

Computer 30 includes an audio storage 32 and a video storage 34 for storing audio and video data provided from microphone arrays 12, 14 and video camera 24, respectively. Computer 30 also includes an indexing engine software module 40 whose operations will be discussed in greater detail below.

Referring to Fig. 2, the hardware for computer 30 used to store and process data and computer instructions is shown. In particular, computer 30 includes a processor 31, a memory storage 33, and a working memory 35, all of which are connected by an interface bus 37. Memory storage 33, typically a disk drive, is used for storing the audio and video data provided from microphone arrays 12, 14 and camera 24, respectively, and thus includes audio storage 32 and video storage 34. In operation, indexing engine software 40 is loaded into working memory 35, typically RAM, from memory storage 33 so that the computer instructions from the

indexing engine can be processed by processor 31. Computer 30 serves as an intermediate storage facility which records, compresses, and combines the audio, video, and indexing information data as the actual conference occurs.

Referring again to Fig. 1, playback system 50 is connected to computer 30 and includes a playback display 52 and a playback server 54, which together allow the recording of the videoconference to be reviewed quickly and accessed at a later time.

Although a more detailed description of the operation is provided below, in general, microphone arrays 12, 14 generate signals, in response to sound generated in the videoconference, which are sent to source locator 22. Source locator 22, in turn, transmits signals representative of the location of a sound source both to a pointing mechanism 26 connected to video camera 24 and to computer 30. These signals are transmitted along lines 27 and 28, respectively. Pointing mechanism 26 includes motors which, in the most general case, control panning, tilting, zooming, and auto-focus functions of the video camera (subsets of these functions can also be used). Further details of pointing mechanism 26 are described in U.S. 5,633,681, incorporated herein by reference. Video camera 24, in response to the signals from source locator 22, is then pointed, by pointing mechanism 26, in the direction of the conference participant who is the current sound source. Images of the conference participant captured by video camera 24 are stored in video storage 34 as video data, along with an indication of the time which has elapsed from the start of the conference.

Simultaneously, the sound picked up by microphone arrays 12, 14 is transmitted to and stored in audio storage 32, also along with the time which has elapsed from the

start of the conference until the beginning of each new sound segment. Thus, the elapsed time is stored with each sound segment in audio storage 32. A new sound segment corresponds to each change, determined by the source locator 22, in the detected location of sound source.

In order to minimize storage requirements, both the audio and video data are stored, in this illustrated embodiment, in a compressed format. If further storage minimization is necessary, only those portions of the videoconference during which speech is detected will be stored, and further, if necessary, the video data, other than the conference participant still images, need not be stored.

Although the embodiment illustrated in Fig. 1 uses one camera, more than one camera can be used to capture the video images of conference participants. This approach is especially useful for cases where one participant may block a camera's view of another participant. Alternatively, a separate camera can be dedicated to recording, for example viewgraphs or whiteboard drawings, shown during the course of a conference.

As noted above, audio storage 32 and video storage 34 are both part of computer 30 and the stored audio and video images are available to both the indexing engine 40 and playback system 50. The latter includes the playback display 52 and the playback server 54 as noted above.

Indexing engine 40 associates the stored video images to the stored sound (segments) based on elapsed time from the start of the conference, and generates a file with indexing information; it indexes compressed audio and video data using a protocol such as, for example, AVI format. For long term storage, audio, video, and indexing information is transmitted from computer 30 to the playback server 54 for

access by users of the system. Playback server 54 can retrieve from its own memory the audio and video data when requested by a user. Playback server 54 stores data from the conference in such a way as to make it quickly available for many users on a computer network. In one embodiment, playback server 54 includes many computers, with a library of multimedia files distributed across the computers. A user can access playback server 54 as well as the information generated by the indexing engine 40 by using GUI 45 with a GUI display 47. Then, the playback display terminal 52 is used to display video data stored in video storage 34 and to play audio data stored in audio storage 32; playback display 52 is also used to display video data and to play audio data stored in playback server 54.

Alternatively, instead of using video images for indexing, an icon is generated based on a still image selected from the continuous video recording. Then, the icon of the conference participant is associated with the audio segment generated by the conference participant. Thus the system builds a database index associating with each identified sound source and its representative icon or image, a sequence of elapsed times and time durations for each instance when the participant was a "sound source". The elapsed times and the durations can be used to access the stored audio and video as described in detail below.

One feature of the invention is to index conference content using the identification of various sound sources and their locations. In the embodiment shown in Fig. 1, the identification and location of sound sources is achieved by the source locator 22 and the two microphone arrays 12, 14. Each microphone array is a PictureTel "LimeLight™" array having four microphones, one microphone positioned at each vertex of an inverted T and at the intersection of the

two linear portions of the "T". In this illustrated embodiment, the inverted T array has a height of 12 inches and a width of 18 inches. Arrays of this type are described in U.S. Patent 5,778,082 by Chu et al., the contents of which are incorporated herein by reference.

In other embodiments, other microphone array position estimation procedures and microphone array configurations, with different structures and techniques of estimating spatial location, can be used to locate a sound source. For example, a microphone can be situated close to each conference participant, and any microphone with a sufficiently loud signal indicates that the particular person associated with that microphone is speaking.

Accurate time-of-arrival difference times of emitted sound in the room are obtained between selected combinations of microphone pairs in each microphone array 12, 14 by the use of a highly modified cross-correlation technique (modified for robustness to room echo and background noise degradation) as described in U.S. 5,778,082. Assuming plane sound waves (the far-field assumption), these pairs of time-differences can be translated by source locator 22 correspondingly into bearing angles from the respective array. The angles provide an estimate of the location of the sound source in three-dimensional space.

In the embodiment shown in Fig. 1, the sound is picked up by a microphone array integrated with the sound localization array, so that the microphone arrays serve double duty as both sound localization and sound pick-up apparatus. However, in other embodiments, one microphone or microphone array can be used for recording while another microphone or microphone array can be used for sound localization.

Although two microphone arrays 12, 14 are shown in use with videoconferencing indexing system 10, only one array is required. In other embodiments, the number and configurations of microphone arrays may vary, for example, from one microphone to many. Using more than one array provides advantages. In particular, while the azimuth and elevation angles provided by each of arrays 12, 14 are highly accurate and are estimated to within a fraction of a degree, range estimates are not nearly as accurate. Even though the range error is higher, however, the information is sufficient for use with pointing mechanism 26.

However, the larger range estimation error of the microphone arrays gives rise to sound source ambiguity problems for a single microphone array. Thus, with reference to Fig. 1, microphone array 12 might view persons 66, 68 as the same person, since their difference in range to microphone array 12 might be less than the range error of array 12. To address this problem, source localization estimates from microphone array 14 could be used by source locator 22 as a second source of information to separate persons 66 and 68, since persons 66 and 68 are separated substantially in azimuth angle from the viewpoint of microphone array 14.

An alternative approach to indexing by sound source location is to use manual camera position commands such as pan/tilt commands and presets to index the meeting. These commands in general may indicate a change in content whereby a change in camera position is indicative of a change in sound source location.

Fig. 3 shows an example of a display 80, viewed on GUI display 47 (Fig. 1), resulting from a videoconference. The following features, included in the display 80, indicate to a user of system 10 exactly who was speaking and when

that person spoke. Horizontal axis 99 is a time scale, representing the actual time during the recorded conference. Pictures of conference participants appear along the vertical axis of display 80. Indexing engine 40 (Fig. 1) selects and extracts from video storage 34 pictures 81, 83, 85 of conference participants 62, 64, 66, on the basis of elapsed time from the start of the conference and the beginning of new sound segments. These pictures represent the conference participant(s) producing the sound segment(s). Pictures 81, 83, 85 are single still frames from a continuous video recording captured by video camera 24 and stored in video storage 34. A key criteria for selection of images for the pictures is the elapsed time from the start of the conference to the beginning of each respective sound segment: the pictures selected for the timeline are the ones which are captured at the same elapsed time as the beginning of each respective sound segment.

Display 80 includes a picture 87, denoting a far-end conference participant. This image, too, is selected by the indexing engine 40. It can be an image of the far-end conference participant, if images from a far-end camera are available. Alternatively, it can be an image of a logo, a photograph, etc., captured by a near-end camera.

Display 80 also includes a block 89 representing, for example, data presented by one of the conference participants at the conference. Data content can be recorded by use of an electronic viewgraph display system (not shown) which provides signals to videoconferencing system 20. Alternatively, a second camera can be used to record slides presented with a conventional viewgraph. The slides, greatly reduced in size, would then form part of display 80.

Associated with each picture 81, 83, 85, 87 and block 89 are line segments representing when sound corresponding to each respective picture occurred. For example, segments 90, 92, 92', and 94 represent the duration of sound produced by three conference participants, e.g. 62, 64, and 66 of Fig. 1. Segment 96 represents sounds produced by a far-end conference participant (not shown in Fig. 1). Segments 97 and 98, on the other hand, show when data content was displayed during the presentation and show a representation of the data content. The segments may be different colors, with different meaning assigned to each color. For example, a blue line could represent a near-end sound source, and a red line could represent a far-end sound source. In essence, the pictures and blocks, together with the segments, provide a series of timelines for each conference participant and presented data block.

In display 80, the content of what each person 62, 64, 66 said is not presented, but this information can, if desired, be filled in after-the-fact by manual annotation, such as a note on the display 80 through the GUI 45 at each speech segment 90, 92, 92', and 94.

A user can view display 80 using GUI 45, GUI display 47, and playback display 52. In particular, the user can click a mouse or other input device (for example, a trackball or cursor control keys on a keyboard) on any point in segments 90, 92, 92', 94, 96, 97, and 98 in the display 80 to access and playback or display that portion of the stored conference file.

A flow diagram of a method 100, according to the invention, is presented in Fig. 4. Method 100 of Fig. 4 is generic to system operation, and could be applied to a wide variety of different microphone array configurations. With

reference also to Figs. 1-3, the operation of system will be described.

In operation, audio is simultaneously acquired from both the far end and the near end of a videoconference. From the far end, audio is continuously acquired for successive preselected durations of time as it is received by videoconferencing system 20 (step 101). Audio received from the far-end videoconferencing system 18 is thus directed to the source locator 22 (step 102). The source locator analyzes the frequency components of far end audio signals. The onset of a new segment is characterized by i) the magnitude of a particular frequency component being greater than the background noise for that frequency and ii) the magnitude of a particular frequency component being greater than the magnitude of the same component acquired during a predetermined number of preceding time frames. If speech is present, an audio segment (e.g., segment 96 in Fig. 3) is begun (step 103) for the timeline corresponding to audio produced by the far-end conference participant(s). An audio segment is continued for the timeline, corresponding to a far-end conference participant, if speech continues to be present at the far-end and there has been no temporal interruption since the beginning of the previously started audio segment.

While the preselected durations of far-end audio are being acquired (step 101) and analyzed, the system simultaneously acquires successive N second durations of audio from microphone arrays 12, 14 (step 104). Because the audio from the far-end site can interfere with near-end detection of audio in the room, the far-end signal received through the microphone arrays is suppressed by the subtraction of a block of N second durations of far-end audio from the acquired near-end audio (step 105). In this

way, false sound localization of the loudspeaker as a "person" (audio source) will not occur. Echo suppression will not affect a signal resulting from two near-end participants speaking simultaneously. In this case, the sound locator locates both participants, locates the stronger of the two, or does nothing.

Echo suppression can be implemented with adaptive filters, or by use of a bandpass filter bank (not shown) with band-by-band gating (setting to zero those bands with significant far-end energy, allowing processing to occur only on bands with far-end energy near the far-end background noise level), as is well-known to those skilled in the art. Methods for achieving both adaptive filtering and echo suppression are described in U.S. 5,305,307 by Chu, the contents of which are incorporated herein by reference.

The detection and location of speech of a near-end source is determined (step 106) using source locator 22 and microphone arrays 12, 14. If speech is detected, then source locator 22 estimates the spatial location of the speech source (step 107). Further details for the manner in which source location is accomplished is described in U.S. 5,778,082. This method involves estimating the time delay between signals arriving at a pair of microphones from a common source. As described in connection with the far-end audio analysis, a near-end speech source is detected if the magnitude of a frequency component is significantly greater than the background noise for that frequency, and if the magnitude of the frequency component is greater than that acquired for that frequency in a predetermined number of preceding time frames. The fulfillment of both conditions signifies the start of a speech segment from a particular speech source. A speech source location is calculated by comparing the time delay of the signals received at the

microphone arrays 12, 14, as determined by source locator 22.

Indexing engine 40 compares the newly derived source location parameters (step 107) to the parameters of previously detected sources (step 108). Due to errors in estimation and small movements of the person speaking, the new source location parameters may differ slightly from previously estimated parameters of the same person. If the difference between location parameters for the new source and old source is small enough, it is assumed that a previously detected source (person) is audible (speaking) again, and the speech segment in his/her timeline is simply extended or reinstated (step 111).

The difference thresholds for the location parameters according to one particular embodiment of the invention are:

1. If the range of both of two sources (previously detected and current) is less than 2 meters, then it is determined that a new source is audible if: the pan angle difference is greater than 12 degrees, or the tilt angle difference is greater than 4 degrees, or the range difference is greater than .5 meters.

2. If the range of either of two sources is greater than 2 meters but less than 3.5 meters, then it is determined that a new source is audible if: the pan angle difference is greater than 9 degrees, or the tilt angle difference is greater than 3 degrees, or the range difference is greater than .75 meters.

3. If the range of either of two sources is greater than 3.5 meters, then it is determined that a new source is audible if:
the pan angle difference is greater than 6 degrees, or the tilt angle difference is greater than 2 degrees, or the range difference is greater than 1 meter.

Video camera 24, according to this embodiment of the invention, is automatically pointed in the response to the determined location, at the current or most recent sound source. Thus, during a meeting, a continuous video recording can be made of each successive speaker. Indexing engine 40, based on correlating the elapsed times for the video images and sound segments, extracts still images from the video for purposes of providing images to be shown on GUI display 47 to allow the user to visually identify the person associated with a timeline (step 109). A new segment of data storage is begun for each new speaker (step 110).

Alternatively, a continuous video recording of the meeting can be sampled after the meeting is over, and still video images, such as pictures 81, 83, and 85 of the participants, can be extracted by the indexing engine 40 from the continuous stored video recording.

Occasionally, a person may change his position during a conference. The method of Fig. 4 treats the new position of the person as a new speaker. By using video pattern recognition and/or speaker audio identification techniques, however, the new speaker can be identified as being one of the old speakers who has moved. When such a positive identification occurs, the new speaker timeline (including, for example, images and sound segments, 85 and 94 in Fig. 3) can be merged with the original timeline for

the speaker. Techniques of video-based tracking are discussed in a co-pending patent application (Serial No. 09/79840, filed May 15, 1998) assigned to the assignee of the present invention, and the contents of which are hereby incorporated by reference. The co-pending application describes the combination of video with audio techniques for autopositioning the camera.

In some cases, more than one conference participant may appear in a still image. The timeline can also indicate any other conference participant who might also appear in the still image (for example, a neighbor sitting in close proximity to the speaker), but is silent at the particular elapsed time, thus giving a comprehensive overview of the sounds produced by all conference participants, as well as helping identify all persons present in the still images.

Conference data can also be indexed for a multipoint conference in which more than two sites engage in a conference together. In this multipoint configuration, microphone arrays at each site can send indexing information for the stream of video/audio/data content from that site to a central computer for storage and display.

Additions, deletions, and other modifications of the described embodiments will be apparent to those practiced in this field and are within the scope of the following claims.

Claims

1. A method for indexing the content of a conference with at least one participant, said method comprising:

recording an audio recording of the conference;

identifying a conference participant producing a sound;

capturing a still image of the identified conference participant;

correlating the still image of the conference participant to at least one audio segment portion of the audio recording, said at least one segment corresponding to the sound produced by the identified conference participant; and

generating a timeline by creating at least one speech-present segment representing the correlated still image and associated at least one audio segment.

2. The method claimed in claim 1, further comprising:

displaying the timeline on a display monitor; and

accessing the timeline displayed on the monitor using a graphical user interface (GUI).

3. The method claimed in claim 2, wherein capturing the still image includes making a video recording of the conference and capturing a video image of the conference participant producing the sound from a segment of the associated video recording of the conference, and further comprising:

using the GUI to select a portion of a specific audio segment for replaying portions of the audio and video recordings on a playback monitor.

4. The method of claim 1, wherein capturing the still image comprises capturing a video image of the conference participant producing the sound from a segment of an associated continuous video recording of the conference.

5. The method of claim 1, further comprising using a video camera to capture the still video image.

6. The method of claim 1 wherein identifying the conference participant is based on identifying the location of the participant.

7. The method of claim 6, wherein identifying the conference participant includes using a microphone array.

8. The method of claim 1, further comprising:

storing time elapsed from a start of the conference with the audio segment and the still image,

wherein the timeline is generated by an indexing engine matching the elapsed time associated with the audio segment and the still image.

9. The method of claim 1, further comprising:

identifying a plurality of conference participants;

capturing a still image of each one of the plurality of conference participants;

storing a time elapsed from a start of the conference indicating the time of the capturing of each still image; and

storing a time elapsed from a start of the conference in association with the audio recording each time a change in audio source location is identified,

wherein generating a timeline includes indicating for each identified conference participant the particular elapsed times from the start of the conference during which the particular participant was speaking, and wherein generating the timeline includes indicating any other conference participant who also appears in the video image and is silent at the particular elapsed time.

10. The method of claim 9, wherein a conference participant has been previously identified and wherein a speech-present segment is added to the timeline for the previously detected conference participant when the participant speaks.

11. The method of claim 10, wherein each identified conference participant is a near-end conference participant.

12. The method of claim 11, wherein identifying each near-end conference participant is based on location.

13. The method of claim 12, wherein a still image of a new near-end conference participant is identified and a new timeline is started for the new near-end conference participant, if the location of the new near-end conference participant is different from previously detected locations of the other identified near-end conference participant.

14. The method of claim 1, wherein the audio source is a far-end loudspeaker transmitting a sound from a far-end speech source, wherein the timeline is a far-end timeline, and wherein generating the far-end timeline includes creating a speech-present segment on the far-end timeline if a far-end speech source is present.

15. The method of claim 14, further comprising:

accumulating a block of far-end loudspeaker microphone array data;

accumulating a block of near-end microphone array data; and

suppressing echo by subtracting accumulated far-end loudspeaker data from accumulated near-end microphone array data.

16. The method of claim 1, further comprising:

capturing a video image of a display presented at the conference; and

generating a timeline for the captured video image of the display.

17. The method of claim 1, wherein the generated timeline is color-coded.

18. A system for indexing the content of a conference with at least one participant, said system comprising:

a sound recording mechanism which records sound created by a conference participant;

at least one source locator for identifying the location of a conference participant, wherein the source locator generates signals corresponding to the location of the conference participant;

a camera assembly including a camera and a camera movement device, which, in response to the signals generated by said source locator, moves the camera to point at the conference participant;

an image capture unit for capturing an image of the conference participant;

an image storage device for storing images captured by said image capture unit;

a processor for associating the image captured by the camera to the sound recorded by the sound recording mechanism and to create a timeline comprising images and indicators of presence of associated sound; and

a graphical user interface which allows access to the stored sound, images, and timeline.

19. The system of claim 18, wherein the sound locator uses at least one microphone array.

20. The system of claim 18, wherein the sound locator uses a plurality of microphones.

21. The system of claim 18, wherein the sound locator comprises a plurality of microphone arrays.

22. A system for indexing the content of a conference with at least one participant, said system comprising:

means for recording an audio recording of the conference;

means for identifying each conference participant producing a sound;

means for capturing a still image of each identified conference participant; and

means for associating the still image of each identified conference participant to at least one audio segment portion of the audio recording corresponding to the sound produced by such conference participant.

23. A method for presenting an audio index database representation of a conference comprising:

generating a plurality of participant timelines, each timeline having at least one speech-present segment representing a correlated still image and at least one associated audio segment;

enabling a user to identify any of the segments representing audio desired; and

playing back the identified segment.

24. A method for indexing the content of a conference with at least one participant substantially as herein described with reference to Figures 1 to 4.

25. A system for indexing the content of a conference with at least one participant constructed and arranged substantially as herein described and shown with reference to Figures 1 to 4.

26. A method for presenting an audio index database representation of a conference substantially as herein described with reference to Figures 1 to 4.

The Patent Office

INVESTOR IN PEOPLE

Application No: GB 9916394.1
Claims searched: 1,18,22

Examiner: Joe McCann
Date of search: 13 January 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): H4F(FEHM, FDX)

Int Cl (Ed.7): G06F(17/60);G11B(27/34);H04N(7/15)

Other: Online: WPI, EPODOC, JAPIO, INSPEC

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
Y	EP 0660249A1	(AT&T CORP) - see abstract	1,18,22
Y	WO 97/01932A1	(AT&T CORP) - see abstract	1,18,22
Y	US 5786814A	(XEROX CORP) - see abstract	1,18,22
Y	US 5729741A	(GOLDEN ENTERPRISES) - see abstract	1,18,22
Y	US 5717869A	(XEROX CORP) - see absract	1,18,22
Y	JP 06-205151A	(FUJI XEROX CO LTD) - see abstract	1,18,22

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.